Finishing half of B1 on Duolingo comparable to five university semesters in reading and listening

Duolingo Research Report DRR-21-03
November 11, 2021 (17 pages)
https://www.duolingo.com/efficacy

Xiangying Jiang*, Haoyu Chen*, Lucy Portnoff*, Erin Gustafson*, Joseph Rollinson*, Luke Plonsky†, and Bozena Pajak*

Abstract

Duolingo is a commercial language-teaching platform that offers free courses on mobile apps and on the web. This study reports the ACTFL listening and reading proficiency levels of adult Duolingo learners who had completed the first seven units of course content in Spanish or French (from the English user interface). The participants (n=340) were learners who had little to no prior proficiency in the target language and used Duolingo as their only learning tool. The participants of the study reached Intermediate Mid/High in reading and Intermediate Low in listening on the ACTFL scale. Their reading and listening scores were comparable with those of university students at the end of the fifth semester of study as reported in Tschirner (2016). The median amount of time learners spent on completing the first seven units was 203 hours. The findings of the study suggest that Duolingo can be an effective tool for foreign language learning at an intermediate level, especially in developing reading and listening skills.

Keywords

Duolingo, efficacy, Spanish, French, reading proficiency, listening proficiency, foreign language

1 Introduction

Duolingo is a language-teaching platform that offers free online courses available on mobile apps and the web. Previous research has shown that the beginning portion of the Spanish and French courses (the first five units) are effective in teaching listening, reading, and speaking skills (Jiang et al., 2021; Jiang, Rollinson, Plonsky, & Pajak, 2020). Specifically, Jiang et al. (2020) showed that Duolingo learners who completed the first five units of Spanish or French on Duolingo achieved proficiency levels in reading and listening comparable to US-based university students who took four semesters of Spanish or French language classes. In addition, Jiang et al. (2021) demonstrated that more than half of Duolingo learners who completed the first five units of Spanish or French on Duolingo met or exceeded the curriculum expectations in speaking skills.

Since more course content has been made available to Duolingo learners of Spanish and French, we aimed to investigate the effectiveness of the new course material. Toward this end, the goals of the present study were (1) to evaluate the listening and reading proficiency levels of Duolingo learners who completed the first seven units in its Spanish or French course, (2) to understand the amount of time it took learners to reach that point in the courses and their in-app behaviors, and (3) to provide a concrete and easily understandable reference point for the Duolingo proficiency levels by benchmarking with US-based university students' language proficiency data.

The Duolingo course units are aligned with the Common European Framework of References (CEFR), an international standard for describing the abilities of language learners at various levels of proficiency. The CEFR divides language proficiency into three broad levels – A (Basic User), B (Independent User), and C (Proficient User), which correspond to the traditional beginner, intermediate, and advanced levels (Council of Europe, 2001). Each broad level is then further divided into two levels, namely, A1 and A2, B1 and B2, and C1 and C2 (see Figure 1).

As shown in Table 1, the current Duolingo Spanish and French courses (from the English user interface) have a total of nine units. Units 1-5 cover the material through the A2 level, while Units 6-9 cover B1-level material. This study, however, only investigated the first seven units, which provide coverage through half of B1. At the time when data collection started, the second half of B1 content was only newly available or not yet fully available to all learners.

Each unit in the Duolingo course structure starts with a unit number in a castle icon and concludes with a checkpoint gate with the number of the completed unit; see Figure 2 for the start of Unit 1 and the end of Unit 7. Each circle in Figure

*Duolingo, Inc.
†Northern Arizona University

Corresponding author:
Xiangying Jiang
Duolingo, Inc. 5900 Penn Ave
Pittsburgh, PA 15206, USA
Email: assessment-study@duolingo.com

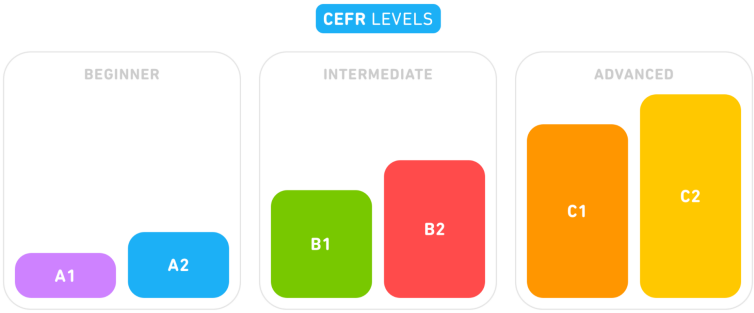

Figure 1. CEFR levels.

Table 1. The Structure of the Duolingo Spanish and French Courses

CEFR level	Course unit	Spanish: # of skills	French: # of skills
Pre-A1	1	9	12
A1	2	29	29
	3	32	25
A2	4	29	28
	5	30	24
B1	6	29	28
	7	28	30
	8	*29*	*27*
	9	*28*	*28*

Note: The current study focuses on learners who complete course content through Unit 7.

2 represents a "skill," which is a collection of lessons on either a communicatively functional topic (such as travel-related vocabulary and expressions, or ordering at a restaurant) or a grammar-focused topic (such as present tense conjugation or pronouns), although note that skills focusing on functional topics also cover new grammatical topics in a less targeted way than grammar-focused skills. There are a total of 164 skills on functional topics and 22 grammar skills in the first seven units of the Spanish course. The first seven units of the French course include 151 skills on functional topics and 25 grammar skills. Each skill on a functional topic includes 5 difficulty levels and each grammar skill has 2 levels. There are 4-5 lessons at each level. Learners are required to complete at least one difficulty level in each skill in a row to unlock the next row, but they can choose whether to complete more levels or not. Learners can attempt to place out of a level by taking a short test; if no more than three exercises are answered incorrectly, the learner is automatically placed in the next level up. Lessons are the primary method of teaching new content on Duolingo, but other modes of learning are available outside of the main course structure. For example, learners can complete generalized practice sessions, which review content they have studied throughout the entire course. For skill-specific practice, learners can return to any skill for which they have completed all difficulty levels in order to refresh their knowledge of a particular functional topic or grammar concept. Another relevant feature is Stories, which provides discourse-level reading and listening comprehension practice, reinforcing and enriching learners' knowledge by situating the unit's content in everyday contexts. Due to the large degree of user autonomy in navigating the platform, there is considerable variation in the types of sessions that learners choose to complete. As a result, there can be substantial variation among individual learners on both the percentage of content they complete before reaching the end of Unit 7 as well as on the total amount of time spent learning.

1.1 The Current Study

The current study aimed to evaluate the effectiveness of Duolingo courses by measuring the listening and reading proficiency levels of learners who had completed the first seven units in the Spanish or French course. The participants had little to no prior proficiency in the target language and used Duolingo as their only learning tool. In particular, the current study investigated the following research questions:

1. What levels of reading and listening proficiency did Duolingo learners achieve upon completing the first

Figure 2. Visual representation of an example Duolingo course structure.

seven units in the Duolingo Spanish or French course? (RQ1)
2. What were the properties of Duolingo learners' in-app activity – in terms of time spent on learning, completing lessons in higher levels, and specific Duolingo features used – before reaching the end of Unit 7? (RQ2)
3. How did Duolingo learners' reading and listening proficiency scores align with an external benchmark? (RQ3)

In the following sections, we describe the method of data collection and report the results of the current study. We end the paper with a brief discussion.

2 Methods

2.1 Participants

The participants of the current study were 208 Spanish learners and 132 French learners on Duolingo who studied these languages from the English user interface. Unlike in Jiang et al. (2020), the participants of the current study were not restricted to learners within the United States. Instead, they were global learners on Duolingo. Because most Duolingo learners are outside the United States, the global sample in this study better represents the Duolingo learner population. The participants learning Spanish were located in 43 countries, with the highest representations from the United States (about 28%), followed by the UK (15%), Canada (6%), and Italy (5%). The participants learning French were located in 39 countries, with the highest representations from the United States (about 35%), followed by the UK (7%), India (6%), and the Netherlands (4%). Overall, 58 countries were represented by the participants in the study.

A combination of in-app data and response to background survey questions was used to select participants who met all the criteria listed below; for more details about the survey, see the Instruments section below. The participants were:

1. learners who reached the end of Unit 7 within the data collection window. End of Unit 7 marks the completion of the first half of B1 course content on Duolingo.
2. learners who self-reported using Duolingo as their only language learning tool. They confirmed that they did not take classes or use other programs or apps during their Duolingo course.
3. learners who had self-reported having no or little prior proficiency in the target language prior to beginning the Duolingo course. In particular, we included only those learners who reported prior proficiency of 0-2 on a 0-10 scale, with 0 representing "I have no knowledge of the language at all," and 10 indicating "I have perfect knowledge of the language." Note that Duolingo collects this information from all learners upon completion of Unit 1 for the purposes of learner analytics and not for course placement.
4. learners aged 18 or older.

Demographic and other background information was also collected through the survey; see Appendix A for specific by-course information. Some general characteristics of the participants are as follows. The average age was 42 for the participants learning Spanish and 38 for those learning French, and there were more male than female participants in both language learner samples. Eighty-seven percent of the participants in both courses reported holding at least a bachelor's degree, with 58% of Spanish participants and 49% of French participants having graduate degrees. Approximately 90% of the participants reported speaking only one language at home before

Table 2. ACTFL Ratings and Numerical Coding

ACTFL levels	Sublevels/ratings	Acronym	Numerical Coding
Novice	Novice Low	NL	1
	Novice Mid	NM	2
	Novice High	NH	3
Intermediate	Intermediate Low	IL	4
	Intermediate Mid	IM	5
	Intermediate High	IH	6
Advanced	Advanced Low	AL	7
	Advanced Mid	AM	8
	Advanced High	AH	9
Superior	Superior	S	10

age 6 (with English being the childhood language for about 40% of the participants), and about 10% of the participants reported speaking more than one language. The fact that about half of the participants who did not speak English at home before age 6 but chose to learn Spanish or French on Duolingo from the English user interface indicated that these participants were bilingual or multilingual. Finally, the participants reported various reasons for learning the language. More than 80% of the participants said that they chose to learn the language for fun/leisure, followed by learning the language for travel (61% in Spanish and 52% in French) and for memory/brain acuity (50% in Spanish and 45% in French).

2.2 Instruments

2.2.1 The Background Survey

The background questionnaire included questions related to participants' language background, reasons for learning the language, level of education, age, and whether they took classes or used other programs/apps during the time they used Duolingo. The latter question confirmed eligibility to satisfy Criterion #2 for participant selection; see Participants above.

2.2.2 ACTFL Listening and Reading Proficiency Tests

We used the ACTFL listening and reading proficiency tests as our main data collection instruments. The ACTFL listening and reading proficiency tests are standardized tests for the global assessment of reading and listening ability (ACTFL, 2013, 2014). They measure how well test-takers spontaneously comprehend the texts and discourse they read or listen to as described in the ACTFL Proficiency Guidelines (ACTFL, 2012). The ACTFL proficiency scale includes four broad levels: Novice, Intermediate, Advanced, and Superior, and each of the first three levels includes three sublevels: low, mid, and high. Altogether, there are ten levels in its proficiency rating scale, from low to high in the order of Novice (low, mid, high), Intermediate (low, mid, high), Advanced (low, mid, high), and Superior. We used Form E of the tests in this project, which targeted proficiency levels between Novice Low and Advanced Low. The tests were administered to each participant online by a remote proctor. Participants were asked to read or listen to 15 passages and answer three multiple-choice questions after each passage. Each test was given an ACTFL rating immediately after the test was submitted. In line with previous research involving ACTFL ratings (e.g. Isbell, Winke, & Gass, 2019; Loewen, Isbell, & Sporn, 2020; Rubio & Hacking, 2019; Tschirner, 2016), we coded each level numerically following a 1-10 point scale. See Table 2 for the mapping between the point scale and each proficiency level.

2.3 Data Collection Procedures

We sent an email soliciting participation in the research study to a random sample of Duolingo learners when they completed Unit 7 in the Spanish or French course, if their self-reported prior proficiency in the language was 0-2. Learners aged 18 and above who were interested in participating completed a background survey within two weeks to verify eligibility and provide additional demographic information. Learners who responded that they had taken classes or used other programs/apps to learn the language during the time they used Duolingo were not invited to participate.

Qualified participants were emailed on a rolling basis and invited to take the ACTFL reading and listening proficiency tests one at a time, with the order of tests (reading, listening) randomized across participants. Qualified participants were first contacted by Duolingo researchers and then received an email from Language Testing International (LTI) with their test ID and instructions about how to schedule a time for the test. The participants had two weeks to finish the test. After the participants finished the first test, we ordered the second test for them and they were again contacted by LTI to take the second test. They went through the same process to schedule and take the test within two weeks. Each participant was paid $100 after completing both tests.

Among a total of 208 Spanish-learner participants, we collected 206 reading and 208 listening scores. Among a total of 132 French-learner participants, we collected 132 reading and 130 listening scores.

2.4 Analyses

Descriptive statistics and correlations were used to answer the first and second research questions on the proficiency outcomes of Duolingo learners and their in-app activity. For the third research question on the comparison of proficiency outcomes between university students and Duolingo learners, t-tests were carried out for each language skill with the R statistical package (R Core Team, 2020).

3 Results

3.1 Proficiency Outcomes of Duolingo Participants

The reading and listening proficiency ratings of Duolingo learners who participated in the current study are visually presented in Figure 3. The distribution of the Spanish reading proficiency ratings was negatively skewed, with 63% of the participants rated at Advanced Low. This indicates a ceiling effect due to the administered test form, which only targeted proficiency levels between Novice Low to Advanced Low. The distributions of the French reading and listening proficiency ratings appeared more normal. Compared with reading proficiency ratings, there were more listening proficiency ratings at the Novice level in both Spanish and French.

Table 3 shows summary data with mean scores and standard deviations calculated by following the numerical coding of the proficiency ratings based on the 1-10 point scale presented in Table 2 above (see Appendix B for participant scores segmented by the most represented countries). On average, Spanish reading scores reached Intermediate High and French reading scores reached Intermediate Mid, while the listening scores of both courses were at Intermediate Low. The participants' reading and listening scores were moderately and significantly correlated, with a Spearman correlation coefficient of 0.64 ($p < .001$) in Spanish and 0.65 ($p < .001$) in French.

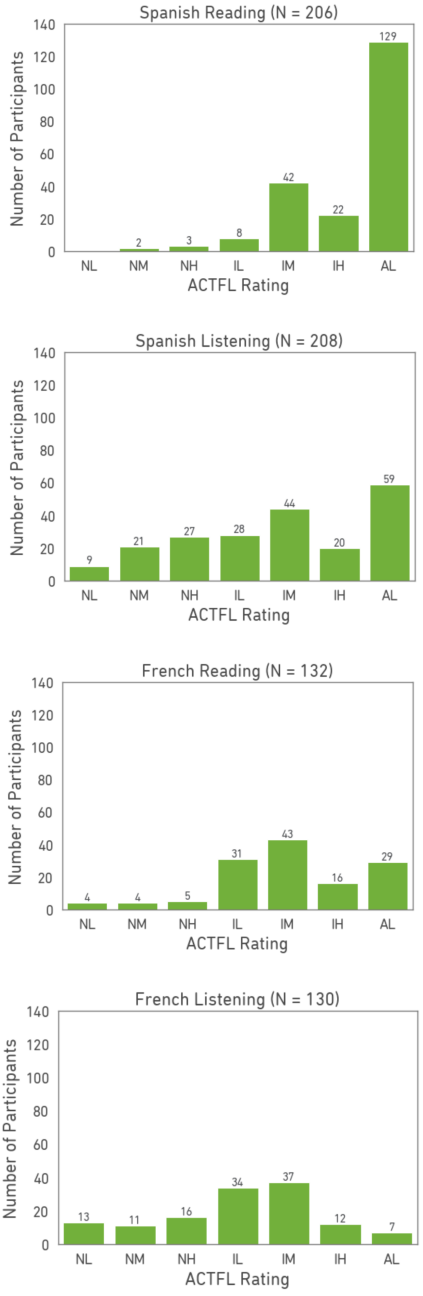

Figure 3. Distribution of ACTFL proficiency ratings of Duolingo study participants (see Table 2 for rating acronyms).

3.2 Duolingo Learners' In-App Behaviors

3.2.1 Time Spent on Learning In order to understand how efficiently learners can learn from the Duolingo Spanish and French courses, we calculated the amount of time Duolingo participants took to reach the end of Unit 7. We computed the total number of hours that the study participants spent in all Duolingo sessions in the target course, as summarized in Table 4 (see Appendix C for calculation details). Due to the lack of a normal distribution (see Figure 4), the median number of hours and interquartile range were reported in addition to the mean and standard deviation. Overall, the participants in the study spent a median of 203 hours working through the first seven units of course content, with participants in Spanish spending a median of 227 hours and those in French a median of 171 hours.

As shown in Figure 4, the number of hours spent by the participants shows a positively-skewed distribution. There was large variation in the amount of time learners spent studying the target languages on Duolingo. This was expected due to the large degree of variation in user behavior resulting from autonomy in

Table 3. Spanish and French Reading and Listening Scores of Duolingo Study Participants

	N	Mean score (SD)	ACTFL rating
Spanish Reading	206	6.25 (1.10)	Intermediate High
Spanish Listening	208	4.79 (1.86)	Intermediate Low
French Reading	132	5.04 (1.47)	Intermediate Mid
French Listening	130	4.04 (1.59)	Intermediate Low

Table 4. Hours Spent in Duolingo Courses by Participants of the Study

Course	Mean (SD)	Median	Min	Max	Interquartile range
Overall (n=340)	249 (192)	203	18	1491	114 - 332
Spanish (n=208)	271 (202)	227	18	1491	129 - 357
French (n=132)	214 (171)	171	28	1106	87 - 283

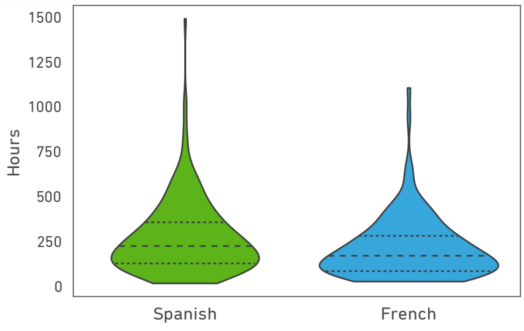

Figure 4. Distribution of hours spent in the target course by study participants, as a symmetric rotated kernel density plot. Each curve shows the full distribution of the data, with dashed lines at the 25th, 50th, and 75th percentiles.

course navigation. For example, learners are only required to complete the first of five difficulty levels in each skill before moving on to the next row. Furthermore, some learners chose to take advantage of additional learning opportunities through generalized and/or skill-specific practice sessions for content review and the Stories feature, among others. Therefore, we observed large between-participant differences in the number of hours spent learning on Duolingo. We did not observe any significant correlation between time spent on learning and proficiency outcomes in French, but found a significant negative correlation for Spanish as shown in Table 5. We outlined possible explanations for this result in the discussion section.

Figure 5 shows the time spent learning (in hours) for learners by ACTFL rating, further illustrating a lack of a clear relationship between time spent on learning and proficiency outcomes. Many of the highest achieving learners spent relatively little time learning, which suggests that the Duolingo app is an efficient learning tool at least for some learners. More research is needed

Table 5. Correlations between Proficiency Outcomes and Time Spent on Learning (TSL)

	Spearman's rho (p-value)
Spanish Reading and TSL	-.26 (p = .0001)
Spanish Listening and TSL	-.31 (p < .0001)
French Reading and TSL	.04 (p > .05)
French Listening and TSL	.015 (p > .05)

to understand the relationship between time spent learning on Duolingo and ACTFL test outcomes. In particular, future research should investigate the characteristics and learning behavior of high-achieving learners.

3.2.2 Duolingo Learners' In-App Activity To better understand the participants' learning behavior in the app, we investigated the various sessions that the participants completed in the 60 days before they were recruited to participate in the study. On the days the participants studied during this 60-day window, Spanish learners completed an average of 7.1 sessions per day and the French learners completed an average of 7.5 sessions per day of study. The boxplot in Figure 6 shows the number of sessions learners completed per day of study.

Due to the large degree of user autonomy in navigating the platform, there is considerable variation in the types of sessions that learners choose to complete. As previously mentioned, learners are required to complete lessons at the first difficulty level in every skill, but can skip levels 2-5. If they want to level up a skill without completing all the lessons, they also have the option of taking a short test to attempt placing out of a level. Figure 7 shows the distribution of Duolingo lessons and place-out tests completed as a percentage of all sessions completed by a given participant in the 60-day analysis window.

The left panel of Figure 7 shows that lessons in the lowest difficulty level (Level 1) comprised the majority of studying sessions for many participants: Level 1 lessons made up at least half of all sessions completed for more than 25% of participants

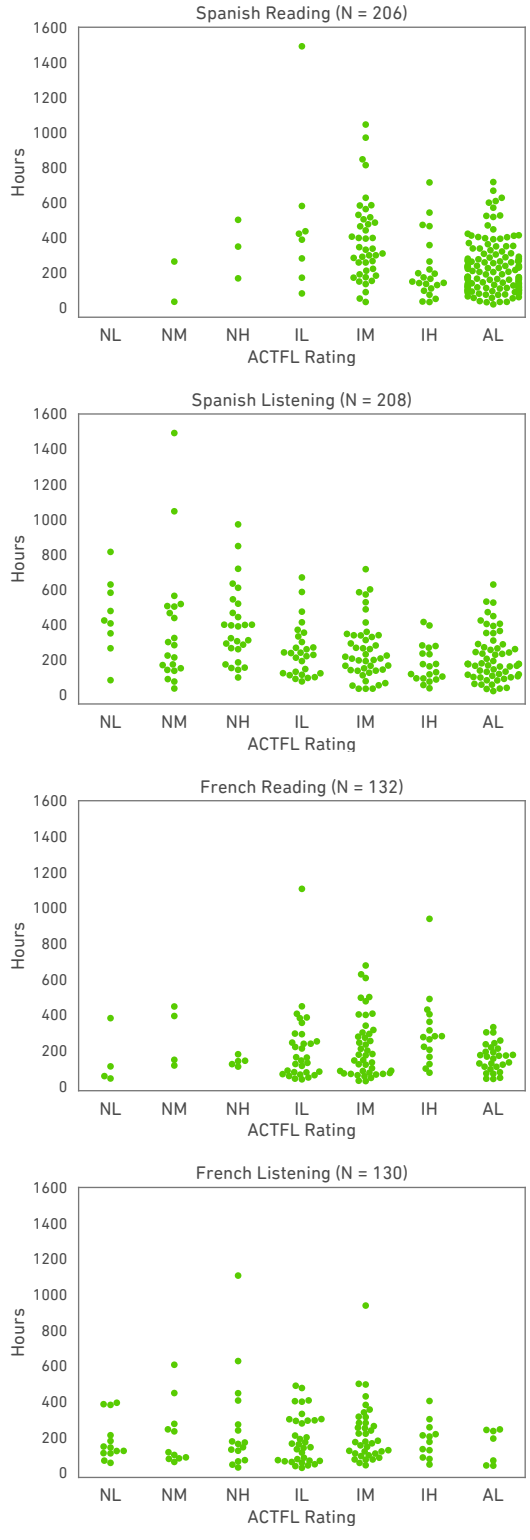

Figure 5. Participants' time spent on learning (in hours) in each ACTFL rating.

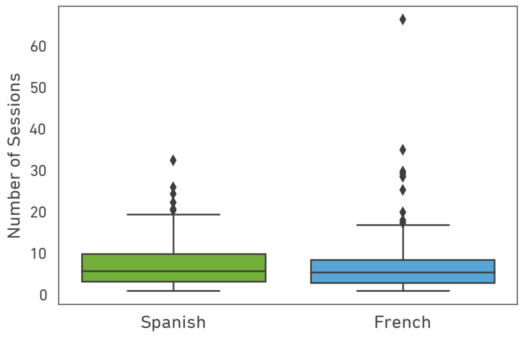

Figure 6. Distribution of the number of sessions completed per learner per active day.

(Spanish: 23.1%; French: 30.3%). We also observe bimodal distributions in Levels 3, 4, and 5, where some participants chose to focus on less difficult levels and others chose to "level up" skills to level 5. The right panel of Figure 7 shows that only 3% of participants attempted to test out of levels for the majority of their sessions, with most learners having little-to-no engagement with this feature.

Some participants also utilized course content outside the standard lessons, such as Stories, skill-specific practice, and generalized practice. As shown in Figure 8, Stories were the most popular non-lesson feature, accounting for a median of 10% (Spanish: 9.4%; French: 10.7%) of participants' total sessions. By contrast, skill-specific and generalized practice were rarely utilized, composing a median of less than 1% of learners' sessions.

3.3 Benchmarking with US-Based University Student Proficiency Data

The third research question of this study involved aligning the reading and listening proficiency scores of Duolingo learners with an external benchmark. In a previous study, Jiang et al. (2020) compared the reading and listening proficiency scores of Duolingo learners who completed the first five course units with fourth-semester university students in the United States. Likewise, in the current study, we compared reading and listening proficiency scores of learners who completed Duolingo Unit 7 with fifth-semester university students. University language programs offer a more traditional setting for foreign language education. While the learner populations at Duolingo and in university classrooms are vastly different, we believe it informative to establish correspondences between learner proficiency outcomes across distinct educational environments. The use of the same standardized tests as measures made this comparison possible.

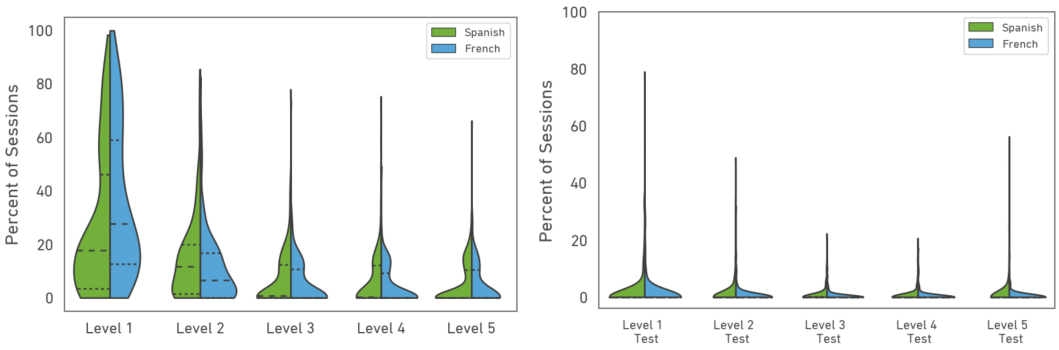

Figure 7. Distribution of Duolingo lessons (left panel) and place-out tests (right panel) completed in the 60 days prior to study recruitment, as a proportion of all sessions completed by a given participant. Distribution for each session type for a given course is shown with a density plot. Tests (right panel) include both successful and unsuccessful attempts.

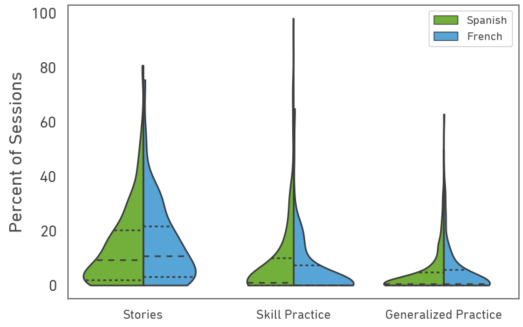

Figure 8. Distribution of Duolingo Stories and practice sessions completed in the 60 days prior to study recruitment, as a proportion of all sessions completed by a given participant. Distribution for each session type for a given course is shown with a density plot.

It is important to note that the curricular foci of the fourth- and fifth-semesters differ substantially in US-based university language programs. In most language programs in the United States, the fourth semester indicates the end of the basic language program (also known as the lower division). The basic language programs provide language instruction to a large number of university students to help them fulfill general education/"liberal studies" or language requirements. The fifth semester marks the beginning of the upper division, which provides more language courses in addition to courses in culture, literature, and linguistics related to the target language. In the upper division, language courses are also more skill-specific; for example, a Spanish class in reading and writing or listening and speaking. The number of students in upper division courses tends to be much smaller than in the lower division. Those who continue language classes beyond the fourth semester may show stronger interest in the target language and may have been higher achieving in earlier semesters. In fact, the majority of the students in upper division language classes are students who study the target language as a major or minor (Winke, Zhang, et al., 2020). Therefore, the variety of learning activities and exposure to the target language on Duolingo and the fifth semester of a university language program can be very different. As a result, it is difficult to estimate the number of semester hours fifth-semester university students spend on language learning. For a better comparison, empirical data would have to be collected in future studies to estimate time spent on language learning among upper-division university students.

The fifth-semester university data we used for external benchmarking were from Tschirner (2016). We chose to use the Spanish and French proficiency data reported in Tschirner (2016) as the source of comparison for two reasons. First, we used two sources for comparison in Jiang et al. (2020): Tschirner (2016) and Rubio and Hacking (2019), but the latter did not report fifth-semester scores. Second, we have access to the Foreign Language Proficiency Test Data from Three American Universities (Winke, Gass, Soneson, Rubio, & Hacking, 2020), where part of Tschirner (2016)'s data came from, but unfortunately, the dataset does not include any data on fifth-semester French students.

Tschirner (2016) reported listening and reading proficiency levels at different milestones of undergraduate study based on data from more than 3,000 participants learning seven languages at 21 institutions across the United States. More concretely, ACTFL listening proficiency tests and reading proficiency tests were administered to first-, second-, third-, and fourth-year students during 2014-2015. Data were collected from learners of French, German, Italian, Japanese, Portuguese, Russian, and Spanish. The main findings were reported based on listening and reading proficiency levels in Spanish and French, which made up 82% of all tests completed. In both languages, there

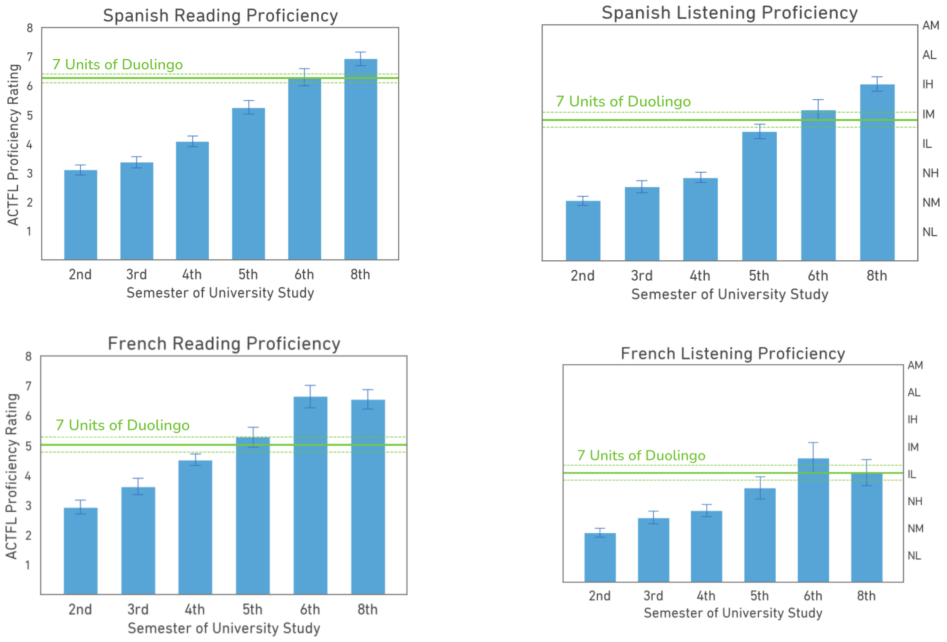

Figure 9. Proficiency scores of Duolingo study participants in relation to US-based university undergraduate students. Error bars represent 95% confidence intervals.

Table 6. Comparison between Duolingo Study Participants and fifth Semester US-Based University Students

Language Skill	Study	N	Mean score (SD)	*t*	*p*	Hedges' *g*	95% CI
Spanish Reading	Duolingo Tschirner (2016)	206 432	6.25 (1.10) 5.26 (2.45)	7.13	<.0001	.47	[0.31, 0.64]
Spanish Listening	Duolingo Tschirner (2016)	208 392	4.79 (1.86) 4.39 (2.44)	2.26	<.05	.18	[0.01, 0.35]
French Reading	Duolingo Tschirner (2016)	132 166	5.04 (1.47) 5.29 (2.20)	-1.18	>.05	-.13	[-0.36, 0.10]
French Listening	Duolingo Tschirner (2016)	130 141	4.04 (1.59) 3.48 (2.51)	2.21	<.05	.26	[0.02, 0.50]

Note: p-values adjusted for multiple comparisons using Benjamini-Hochberg method with alpha=0.01.

was a steady increase in proficiency levels over the semesters in both listening and reading, but listening proficiency levels were substantially lower than reading levels. By the end of the fifth semester, on average, students reached Intermediate Mid in reading proficiency (with a mean score of 5.26 in Spanish and 5.29 in French), and their listening proficiency was Intermediate Low in Spanish (with a mean score of 4.39) and Novice High in French (with a mean score of 3.48 in French). See see Figure 9 for the Spanish and French scores across semesters.

Table 6 summarizes the scores of Duolingo study participants and those of university students at the end of the fifth semester in Tschirner (2016). To assess whether there were significant differences between Duolingo learners and fifth-semester university students, four separate Welch's two-sample t-tests were carried out using the R statistical package (R Core Team, 2020). We conducted comparisons on the summary data (i.e., counts, means, and standard deviations) available to us in Tschirner (2016). Because we conducted four separate t-tests, we adjusted the p-value using the Benjamini-Hochberg method with alpha = 0.01. The t-test results are also included in Table 6. Among the four pairs of comparisons, Duolingo participants scored significantly higher than those of the fifth-semester university students in Spanish reading (t = 7.13, p < .0001, g = .47), Spanish listening (t = 2.26, p < .05, g = .18), and French listening (t = 2.21, p < .05, g = .26), although the effect sizes were small. No significant difference was found between Duolingo study participants and fifth semester university students in French reading.

Since no significant difference was observed between the Duolingo study participants and the fifth-semester university students in French reading, we ran an additional t-test between the Duolingo study participants and the fourth-semester university students. Tschirner (2016) reported French reading data of 215 fourth-semester university students with a mean score of 4.52 and a standard deviation of 1.45. The results indicated that the Duolingo study participants did significantly better than the fourth-semester university students in French reading (t = 3.20, p < .05, g = .36), thus providing additional evidence that they were likely at a level comparable to fifth-semester students.

4 Discussion and Conclusion

In this study, we assessed the reading and listening proficiency of Duolingo learners who had completed the first seven units of the Spanish or French course on Duolingo, analyzed their in-app activities, and compared their proficiency scores to those of fifth-semester US-based university students on the same measures. The main findings are summarized and discussed below.

4.1 Reading and Listening Proficiency Outcomes after Unit 7

ACTFL proficiency scores indicated that Duolingo learners who had completed the first seven units of the Duolingo Spanish or French course reached, on average, Intermediate Mid or Intermediate High in reading and Intermediate Low in listening. Although the participants' reading and listening scores were moderately correlated, the listening proficiency of Duolingo learners was significantly lower compared to reading proficiency, which replicated the findings of Jiang et al. (2020) for Duolingo learners and Tschirner (2016) for university students. Although both listening comprehension and reading comprehension are receptive skills, the comprehension processes have been found to be mostly modality-specific (Wolf, Muijselaar, Boonstra, & Bree, 2019). For learners at early stages of language learning, listening comprehension demands a higher level of attention, exerts a heavier load on working memory, and requires the ability for speedy decoding and processing of transient audio input (see, e.g., Bloomfield et al., 2010; Wallace, 2020). In contrast, learners' decoding process in reading is facilitated by the availability of visually-presented text (Spoden, Fleischer, & Leucht, 2020; Vandergrift & Baker, 2015). As a result, listening comprehension is often more challenging than reading comprehension for second language learners and takes longer to develop. Tschirner (2016) also attributed students' lower listening proficiency to insufficient attention to auditory input and exercises in classroom instruction and called for more emphasis on listening development in instructional practices.

Language learning apps are thought to be good for developing decontextualized linguistic knowledge, and Duolingo is considered one example of such apps (Krashen, 2014). Although most Duolingo lessons focus on vocabulary and grammar at the sentence level, the findings of this study demonstrate that learners were able to transfer discrete linguistic knowledge to integrative tasks such as reading and listening comprehension, as well as speaking (Jiang et al., 2021). Some Duolingo learners might have also benefited from the Stories feature, which provides discourse-level reading, listening, and speaking practice and is one of the most popular features on Duolingo. Loewen et al. (2020) observed similar evidence of discrete linguistic knowledge being transferred to speaking tasks, and proposed that the field of second language acquisition should "abandon earlier characterizations of language learning apps as merely 'mechanical practice of selected and graded grammatical phenomena…in the form of drills' " by citing Heift and Vyatkina (2017) and called for the field to "recognize the pedagogical potential of widely used modern apps" (p. 19). The authors of this study concur with Loewen et al. on this proposal.

4.2 Duolingo Participants' In-App Activities

In-app analysis demonstrated that the median amount of time that the participants took to complete the first seven units was 203 hours (227 hours for Spanish learners and 171 hours for French learners). Data on in-app behavior in the 60 days prior to study recruitment showed that the majority of the lessons completed were at Level 1, which is the lowest level of five and the only one required to progress through the course. Learners also took advantage of other features such as Stories, skill-specific practice, and generalized practice. The self-directed nature of the Duolingo learning platform contributes to this variation and complicates our interpretation of the relationship between app behavior and learning outcomes; for example, we observed bimodal "leveling up" behavior, where some learners chose to regularly complete more difficult lessons at higher levels while others rarely did. Spearman correlations between time spent on learning and proficiency were negligible and non-significant for French reading and listening and significantly negative for Spanish reading and listening, but interpretation is complicated by several confounding factors. Learners with greater time spent on learning may be spending time in particular session types that teach the material less efficiently, or they may be struggling with course content and taking longer to complete sessions as a result. On the other hand, learners with less time spent on learning may have successfully tested out of a large amount of review-focused content and therefore reached the end of Unit 7 more quickly. In other words, it is possible for learners to make proficiency gains in a relatively short amount of time. Future studies could address these issues and provide stronger signals about this relationship by using a pre- and post-test design, which allows for more control over the time spent learning during the course of the study.

At the same time, the degree of learner autonomy afforded by the Duolingo program can also be considered a strength in the

design of the study in terms of the ecological validity that it provides. Participants of the study were not required to study a certain amount of time per day or per week, or to study a minimal total amount of time to be included in the study. What is captured in the study is learners' intrinsic preferences for how to use the app to address their own needs and interests. The respect for individual learning preferences seems to have successfully motivated learners and facilitated their proficiency development.

4.3 Benchmarking with US-Based University Student Proficiency Data

In comparing listening and reading proficiency between Duolingo learners and university students in language classes, the results indicated that when Duolingo Spanish and French learners reached the end of Unit 7 on Duolingo their proficiency scores in Spanish reading, Spanish listening, and French listening were significantly higher than fifth-semester university students as reported in Tschirner (2016), while their French reading proficiency was not significantly different from fifth-semester university students (but significantly higher than fourth-semester university students). The findings of the current study, together with the findings of Jiang et al. (2020), provide evidence that Duolingo courses are effective for developing reading and listening skills.

The availability of the university proficiency data in Tschirner (2016) made the comparison of Duolingo learners and US university students possible; however, this comparison should not be interpreted as competition between online language learning apps and university language programs. The aim of comparing learning outcomes from the two contexts is, rather, as a means to benchmark the progress made by Duolingo learners relative to a more familiar and traditional setting. It is also important to note some major differences between the participants of the study and the university student sample. First, the university sample were mostly full-time students from a more homogeneous age range of approximately 20-22, while the Duolingo study participants were mostly post-university adults with an average age of 38 (French) and 42 (Spanish). Second, as reported in Winke, Zhang, et al. (2020), the fifth-semester university student sample likely included students with multiple years of K-12 language learning experience or heritage status in the target language, while the Duolingo study participants were restricted to those who had a prior self-reported proficiency of 0-2 (on a scale of 0-10, from no knowledge to to full proficiency) and used Duolingo as the only learning tool. Third, the curricula in the two settings were also very different. In some sense, the Duolingo curriculum is more similar to the lower division university language curriculum, in which enrollment in a language course requires 3-4 hours of coursework per week. Due to substantial changes from lower to upper division language curriculum at US universities, students in the upper division (third and fourth year) language programs may have a lot more exposure to the target language compared to students in the lower division (1st and 2nd year). Fourth, the motivations for language learning also differed. The university students enrolled in fifth-semester language courses were mostly learning the target language as a major or minor for the purpose of an academic degree, while the majority of the Duolingo study participants were learning the language on Duolingo for fun or leisure, for travel, and for memory or brain acuity.

4.4 Limitations and Directions for Future Research

The current study tested learners when they reached the end of Unit 7 independently. For future research, a pre- and post-test design will allow more control of learning time and participant factors that were self-reported in the present study, including prior proficiency, exposure to the target language outside of Duolingo, and the exclusion of other learning tools. This study focused on listening and reading proficiency, which are both receptive skills. Learners were not assessed in speaking (as in Jiang et al., 2021) or writing. In subsequent studies, Duolingo's effectiveness in developing learners' productive skills when they reach the end of Unit 7 will be evaluated as well. Doing so will provide a better understanding of whether and to what extent Duolingo learners' success in receptive skills can also be observed in productive skills.

The findings of the current study do not represent the overall effectiveness of Duolingo or university language courses, so they should not be overgeneralized. Participants of the current study were only compared on reading and listening skills, while teaching effectiveness can be reflected in other skills and abilities. The availability of the university proficiency data in Tschirner (2016) made this comparison possible; however, the comparison between Duolingo learners and university students should not be interpreted as competition between online language learning apps and university language programs. The aim in comparing learning outcomes from the two contexts is, rather, as a means to benchmark the progress made by Duolingo learners relative to a more familiar and traditional setting.

4.5 Pedagogical Implications

The results of the current study indicate that Duolingo can help self-directed learners develop reading and listening proficiency. The proficiency outcomes were comparable to those of US-based university students in their fifth semester of upper-division language programs. Although Duolingo courses mostly teach vocabulary and grammar at the sentence level (with some longer-form content available in the form of short stories and podcasts), the results of this study also suggest that the seemingly discrete vocabulary and grammar knowledge can be applied to integrative tasks such as listening and reading comprehension.

The findings suggest that use of Duolingo can lead to substantial proficiency development as a tool for self-directed study. However, the usage data from the present study did not reveal

a significant positive relationship between learners' total hours spent using the app and their reading and listening scores. The lack of this relationship in the data might be due to vast variability in the time (hours) and intensity of learning that participants took to complete the first seven units of their course. Consequently, it would be premature to make any suggestions regarding when and how the app might be used to maximize its efficiency. However, we plan to address this question in a future study.

In addition to self-directed learners, classroom teachers have used Duolingo to their advantage and to the benefit of their students (Munday, 2016, 2017), suggesting that the app is also a useful tool to complement other types of language instruction. For instance, if vocabulary and grammar practice can be largely done by students as homework using apps such as Duolingo, more class time can be directed toward the teaching of culture and other communicative skills.

4.6 Conclusion

This study evaluated the reading and listening proficiency outcomes of Duolingo learners who had little to no prior knowledge of the target language and used Duolingo as their only learning tool. The findings demonstrated that, on average, learners who finished the first seven units of the Duolingo Spanish or French course reached ACTFL Intermediate level in both reading and listening proficiency. The proficiency scores of Duolingo learners were comparable with the proficiency outcomes of students at the end of their fifth semester in upper-division US-based university language programs. In conducting this study, we hope to have shed light on the potential effectiveness and comparability of Duolingo, as measured through standardized tests, to more traditional settings. Future studies will continue to build on our findings at other levels of study, in other linguistic domains, and in other target languages.

Note

The original title of this paper was "Seven units of Duolingo courses comparable to 5 university semesters in reading and listening." We revised the title of this paper because Duolingo made changes to its home screen design and its labeling system in 2022. This study was conducted when learners were on the previous version. Since Duolingo's CEFR-aligned course content remains the same, the new title of the paper refers to the CEFR level of content completed. Half of the B1 content completed in this study was previously covered in seven "units", and is now covered in five "sections" in the current Duolingo courses.

Author Biographies

Xiangying Jiang is a lead learning scientist and works on learning assessment at Duolingo. She has a Ph.D. in Applied Linguistics and was Associate Professor of TESOL at West Virginia University before joining Duolingo.

Haoyu Chen is currently a software engineer at Duolingo. She graduated from University of Pennsylvania with a Master's degree in Data Science.

Lucy Portnoff is a data scientist focusing on learning assessment at Duolingo. She graduated from the University of California, Berkeley with an undergraduate degree in Mathematics.

Erin Gustafson holds a Ph.D. in Linguistics (Northwestern University, 2016). Prior to joining Duolingo in 2017, Erin was a post-doctoral fellow at the Northwestern University Medical School, where her research focused on machine learning and natural language processing applications in the medical domain. Her graduate research focused on bilingualism and psycholinguistics. She is currently Lead Data Scientist at Duolingo and co-leads a team focused on learning assessment.

Joseph Rollinson is currently a staff software engineer at Duolingo, where he co-leads teams focused on learning assessment and learning infrastructure. He graduated from Carnegie Mellon University with undergraduate degrees in Computer Science and Philosophy. As an undergraduate, he performed research in intelligent tutoring systems.

Luke Plonsky is Associate Professor of Applied Linguistics at Northern Arizona University. His work, focusing primarily on second-language acquisition and research methods, has appeared in over 80 articles, book chapters, and books. Luke is Associate Editor of *Studies in Second Language Acquisition*, Managing Editor of *Foreign Language Annals*, and Co-Director of the IRIS Database.

Bozena Pajak holds a Ph.D. in Linguistics (University of California, San Diego, 2012). Before joining Duolingo in 2015, she was a Research Associate and a Lecturer in Linguistics at Northwestern University. Her research focused primarily on the acquisition of additional languages in adulthood. She is currently the Director of Learning and Curriculum at Duolingo, where she co-leads the company's Learning Area.

5 References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Retrieved from https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines2012.pdf

ACTFL. (2013). *ACTFL reading proficiency test (RPT). Familiarization manual and ACTFL proficiency guidelines 2012 – reading*. Retrieved from https://www.languagetesting.com/pub/media/wysiwyg/manuals/ACTFL_FamManual_Reading_2019.pdf

ACTFL. (2014). *ACTFL listening proficiency test (LPT). Familiarization manual and ACTFL proficiency guidelines 2012 – listening*. Retrieved from https://www.languagetesting.com/pub/media/wysiwyg/manuals/ACTFL_FamManual_Listening_2019.pdf

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*. MARYLAND UNIV COLLEGE PARK.

Council of Europe. (2001). *Common european framework of references for languages: Learning, teaching, assessment*. Retrieved from https://rm.coe.int/1680459f97

Isbell, D. R., Winke, P., & Gass, S. M. (2019). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*, *36*(3), 439–465.

Jiang, X., Rollinson, J., Chen, H., Reuveni, B., Gustafson, E., Plonsky, L., & Pajak, B. (2021). *How well does duolingo teach speaking skills?* Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/duolingo-speaking-whitepaper.pdf

Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). *Duolingo efficacy study: Beginning-level courses equivalent to four university semesters*. Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/duolingo-efficacy-whitepaper.pdf

Krashen, S. (2014). Does duolingo "trump" university-level language learning. *International Journal of Foreign Language Teaching*, *9*(1), 13–15.

Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, *53*(2), 209–233.

Munday, P. (2016). The case for using duolingo as part of the language classroom experience. *RIED: Revista Iberoamericana de Educación a Distancia*, *19*(1), 83–101.

Munday, P. (2017). Duolingo. Gamified learning through translation. *Journal of Spanish Language Teaching*, *4*(2), 194–198.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rubio, F., & Hacking, J. F. (2019). Proficiency vs. Performance: What do the tests show? In *Foreign language proficiency in higher education* (pp. 137–152). Springer.

Spoden, C., Fleischer, J., & Leucht, M. (2020). Converging development of english as foreign language listening and reading comprehension skills in german upper secondary schools. *Frontiers in Psychology*, *11*, 1116.

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals*, *49*(2), 201–223.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416.

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*. https://doi.org/10.1111/lang.12424

Winke, P., Gass, S. M., Soneson, D., Rubio, F., & Hacking, J. F. (2020). *Foreign language proficiency test data from three american universities, [united states], 2014-2017*. Inter-University Consortium for Political; Social Research. https://doi.org/10.3886/ICPSR37499.V1

Winke, P., Zhang, X., Rubio, F., Gass, S., Sonenson, D., & Hacking, J. (2020). The proficiency profiles of language students: Implications for programs. *Second Language Research & Practice*, *1*(1), 25–64. Retrieved from http://hdl.handle.net/10125/69840

Wolf, M. C., Muijselaar, M. M. C., Boonstra, A. M., & Bree, E. H. de. (2019). The relationship between reading and listening comprehension: Shared and modality-specific components. *Reading and Writing*, *49*, 1747–1767.

A Appendix

Table 7. Characteristics of the Participants

Characteristic	Spanish (N = 208)	French (N = 132)
Age		
Mean (SD)	42 (15)	38 (13)
Prefer not to answer	1	0
Home language before age 6		
Only English	40%	38%
Only one language, but not English or the target language assessed in the study	49%	52%
More than one language, but not the target language assessed in the study	11%	10%
Heritage speaker of target language	<1%	0%
Highest level of education		
Bachelor's degree	29%	38%
Graduate degree	58%	49%
Master's degree	44%	39%
Doctoral degree	14%	10%
Other	13%	13%
Reasons for learning the language		
For fun / leisure	82%	83%
For travel	61%	52%
For memory / brain acuity	50%	45%
For job-related purposes	19%	24%
For social purposes	28%	30%
For school	4%	4%
Other	9%	8%
Gender		
Female	46%	37%
Male	53%	62%
Prefer not to answer	1%	1%

B Appendix

Table 8. Proficiency Scores of Participants Segmented by Most Represented Countries

Course	Country	Reading		Listening	
		N	Mean score (SD)	N	Mean score (SD)
Spanish	US	58	5.78 (1.41)	59	3.64 (1.80)
	UK	32	6.28 (1.02)	32	4.50 (1.72)
	CA	12	6.08 (1.0)	13	4.00 (1.87)
	IT	10	7.00 (0.0)	10	6.80 (0.42)
French	US	47	5.09 (1.30)	46	4.00 (1.49)
	UK	9	5.22 (1.56)	9	3.78 (1.30)
	IN	8	3.38 (1.77)	8	2.25 (1.16)
	NL	5	5.40 (1.14)	5	5.40 (1.52)

C Appendix

Calculation of Time Spent on Learning

In this study, we calculated the total time to reach the end of Unit 7 by summing the time spent on learning for each session ever completed by the learner in the target course.

Starting in August 2019, Duolingo calculated time spent on learning for each session by storing and summing the number of seconds spent completing each exercise within a session, with a maximum value of 60 seconds per exercise. The maximum is enforced to exclude non-learning time from the calculation, such as the learner quitting the app mid-session and returning much later; this occurs very rarely and is unlikely to impact the final result.

However, many participants in the study completed sessions before August 2019, when exercise-level time data was not stored. Therefore, we used an alternative approach where we measured the wall-clock time spent in sessions with a maximum value of 10 minutes per session. We then multiplied this value by a constant (determined by linear regression) to be consistent with the exercise-level method.

To validate the wall-clock method, we re-calculated time spent on learning for all sessions completed after August 2019 using this method. We observed a Pearson correlation of 0.974 and a 7.3% average difference between time spent learning from the exercise-level vs. wall-clock methods.